УДК 511

ПРИЕМУЩЕСТВА НЕПАРАМЕТРИЧЕСКИХ СТАТИСТИЧЕСКИХ МЕТОДОВ

Стребков Е.В., кандидат физико-математических наук, доцент,
Казанский Федеральный Университет, г. Казань
str9050258629@yandex.ru

Пашкин А.В., студент,
Казанский Федеральный Университет, г. Казань

Симаков Н.Е., студент,
Казанский Федеральный Университет, г. Казань

Журавлева М.И., магистрант,
Казанский Федеральный Университет, г. Казань

Аннотация. В данной статье рассматриваются рекомендации по применению ранговых методов корреляционного и факторного анализа.

Ключевые слова: аналитическая статистика; корреляционный анализ, факторный анализ; ранговые методы.

THE ADVANTAGES OF NONPARAMETRIC STATISTICAL PRINCIPLES

E.V. Strebkov, PhD, associate professor,
Kazan Federal University, Kazan
str9050258629@yandex.ru
A.V. Pashkin, student,
Kazan Federal University, Kazan
N.E. Simakov, student,
Kazan Federal University, Kazan
M.I. Zhuravliova, master,
Kazan Federal University, Kazan

Abstract. This paper reviews recommendations on efficient application of ranking methods of correlation and factor analysis.

Keywords: analytical statistics; correlation analysis, factor analysis; ranking methods.

Необходимой частью подготовки современных специалистов является изучение методов математической статистики, которые востребованы для многих специальностей, например, физико-математического, информационного, медико-биологического, социально-экономического, психолого-педагогического профилей.

Многообразие и особенности применения статистических методов обуславливают затруднение при их изучении. Поэтому актуальна задача выделения достаточно универсальных и эффективных статистических методов, доступных для применения широким кругом специальностей. В данной статье рассматривается преимущества непараметрических (ранговых) методов для корреляционного и факторного анализа.

В отличие от параметрических методов аналитической статистики непараметрические методы являются более универсальными, т.к. применимы для количественных и качественных признаков без ограничений на законы распределения изучаемых признаков и не требуют сложных вычислений выборочных параметров. Суть ранговых методов состоит в анализе отношений «больше - меньше» между реальными показателями изучаемых признаков. Эффективность непараметрических методов проиллюстрируем на примерах из корреляционного и факторного анализа.

Одним из востребованных является рангово-бисериальный коэффициент корреляции для задач, когда один признак измеряется в дихотомической шкале (признак X), а другой в ранговой шкале (признак Y), применение которого продемонстрируем на примере 1 [2].

Таблица 1

X пол	1	0	1	1	0	1	0	0	1
Y = IQ	101	108	86	91	105	78	93	102	103
Ранги	5	9	2	3	8	1	4	6	7

Пример 1. Исследуется возможность гендерного различия в показателях интеллекта (коэффициент умственных способностей IQ) на примере подростков разного пола. Результаты обследования приведены в Таблице 1.

В Таблице 1 дихотомический признак X = {пол} принимает значения 1 для юношей и 0 для девушек, объем выборки $n = 9$,признак Y = {значение коэффициента IQ} является количественным, значения которого проранжированы в порядке возрастания.

Рассмотрим поэтапный алгоритм с параллельными вычислениями для примера 1.

Этап 1. Определяются средние ранги:

1) для юношей при $X = 1$ $\overline{x_1} = (1 + 2 + 3 + 5 + 7) / 5 = 3{,}6;$
2) для девушек при $X = 0$ $\overline{x_0} = (4 + 6 + 8 + 9) / 4 = 6{,}75.$

Этап 2. Вычисляется выборочное значение рангово-бисериального коэффициента корреляции:

$\bar{r} = \frac{2*(\overline{x_1}-\overline{x_0})}{n} = \frac{2*(3{,}6-6{,}75)}{9} = -0{,}7.$

Этап 3.С целью проверки значимости коэффициента корреляции рассматриваются гипотезы:
нулевая гипотеза $H_0 = \{\text{коэффициент } \bar{r} \text{ значимо не отличается от нуля}\}$;
альтернативная гипотеза $H_0 = \{\text{коэффициент } \bar{r} \text{ значимо отличается от нуля}\}$.

Этап 4. Вычисляется фактическое значение критерия:

$T = \overline{|r|}\sqrt{\frac{n-2}{1-(\bar{r})^2}} = 0{,}7\sqrt{\frac{7}{1-(0{,}7)^2}} = 2{,}59.$

Коэффициент $\bar{r}$ изменяется в диапазоне от -1 до +1, его знак для интерпретации результатов не имеет значение в силу равноправия значений $X = 0$ и $X = 1$.

Этап 5. Исследователем задается уровень значимости q, т.е. вероятность отклонения гипотезы H_0 при ее справедливости. Согласно уровня значимости q и числу степеней свободы $k = n - 2$ по таблице критических значений критерия Стьюдента находят $T_{\text{кр}}$ [1]. Если $T > T_{\text{кр}}$, то принимается гипотеза H_1.

Для примера 1 при $q = 0{,}05$ и $k = n - 2 = 7$ $T_{\text{кр}} = 2{,}36$. Таким образом, $T = 2{,}59 > T_{\text{кр}} = 2{,}36$ и принимается H_1 о значимом отличии рангово-бисериального коэффициента корреляции $\bar{r} = 0{,}7$ от нуля, т.е. на данной выборке подростков выявлено значимое гендерное различие по показателю коэффициента IQ.

Далее рассматривается особенности однофакторного дисперсионного анализа, который применяется, чтобы установить оказывает ли на изучаемый (результирующий) признак X существенное влияние на некоторый качественный фактор F, имеющий несколько уровней.

Применение обычного (параметрического) дисперсионного анализа имеет существенные ограничения [1]:

1) при каждом уровне фактора F изучаемый признак X должен иметь нормальный закон распределения с постоянной для различных уровней генеральной дисперсией;
2) необходимость достаточно трудоемких вычислений выборочных дисперсий.

Преимущества рангового подхода проиллюстрируем на примере однофакторного непараметрического метода на основе критерия Краскала – Уоллеса [2].

Пример 2. Анализируется влияние фактора F= {величина торговой площади в квадратных метрах} с тремя k = 3 уровнями F_1= {площадь до 100 м^2}, F_2= {площадь от 100 м^2 до 150 м^2}, F_3= {площадь более 150 м^2} на показатели результирующего признака X = {количество единиц реализованного товара в течении недели} по 3 видам товаром. Результаты обследования приведены в

Таблице 1, где x_{ij} – значение признака X для уровня F_i фактора F и j-того вида товара, n_i и n – суммарное количества товаров соответственно для уровня F_i и всего фактора F, т.е. $n = n_1 + n_2 + n_3 = 3 + 3 + 3 = 9$.

Рассмотрим поэтапный алгоритм метода с параллельными вычислениями для примера 2.

Этап 1. Формулируется основная гипотеза H_0= {расхождение наблюдений для различных уровней фактора F обусловлено случайными причинами}, альтернативные гипотезы H_1 могут быть произвольными.

Таблица 2

Номер товара	Уровень F_1	Уровень F_2	Уровень F_3
1	$x_{11} = 105$	$x_{21} = 97$	$x_{31} = 120$
2	$x_{12} = 162$	$x_{22} = 171$	$x_{32} = 183$
3	$x_{13} = 194$	$x_{23} = 206$	$x_{33} = 192$
n_i	$n_1 = 3$	$n_2 = 3$	$n_3 = 3$

Этап 2. В Таблице 2 заменим наблюдения x_{ij} их соответствующими рангами r_{ij}, упорядочивая всю совокупность наблюдений в порядке возрастания. В результате получим Таблицу 3, где среднее всех рангов R =(n+1)/2= 5 и средние ранги для уровня F_i равны $R_i = (r_{i1} + r_{i2} + r_{i3})/n_i$.

Таблица 3

Номер товара	Уровень F_1	Уровень F_2	Уровень F_3
1	$r_{11} = 2$	$r_{21} = 1$	$r_{31} = 3$
2	$r_{12} = 4$	$r_{22} = 5$	$r_{32} = 6$
3	$r_{13} = 8$	$r_{23} = 9$	$r_{33} = 7$
R_i	$R_1 = 4{,}66$	$R_2 = 5$	$R_3 = 5{,}33$

Этап 3. Вычисляется наблюдаемое значение критерия

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{3} n_i \left(R_i - \frac{n+1}{2}\right)^2 = \frac{12}{9*10} \sum_{i=1}^{3} 3 * (R_i - 5)^2 = 0{,}09 .$$

Этап 4. Задается уровень значимости q, т.е. вероятность отклонения гипотезы H_0 при ее справедливости. Согласно уровню значимости q и числу степеней свободы $(\mathrm{k} - 1)$ по таблице критических значений распределения хи-квадрат [1] выбираем критическое значение $\chi^2_{\text{кр}}$, k – число уровней F.

Суть факторного метода состоит в сравнении общего среднего ранга R со средними рангами R_i по факторам F_i. При $H > x^2_{\text{кр}}$ гипотеза H_0 отклоняется на уровне значимости q, т.е. влияние фактора F считается значимым.

Для примера 2 при $q = 0{,}05$ и числа степеней свободы $(k - 1) = 2$ критическое значение $x^2_{\text{кр}} = 6{,}0$ и наблюдаемое значение $H = 0{,}09$. Следовательно, $H = 0{,}09 < x^2_{\text{кр}} = 6{,}0$ и гипотеза H_0 принимается, таким образом, фактор $F = \{\text{величина торговой площади}\}$ не оказывает влияние на объем продаж.

По сравнению с параметрическими методами статистики непараметрические (ранговые) методы обладают существенными преимуществами:

1) применимы для значительного числа классов прикладных задач из различных областей знаний;
2) являются универсальными для анализа количественных и качественных признаков;
3) не ограничены жесткими требованиями о законе распределения изучаемых признаков;
4) не опирается на углубленные знания по теории вероятностей о свойствах случайных величин;

5) обладают наглядностью и простотой алгоритмов реализации;
6) не требуют трудоемких вычислений;

Использование непараметрических методов при обучении аналитической статистике способствуют эффективному формированию у учащихся необходимых компетенций для анализа прикладных задач, актуальных для соответствующих специальностей.

Литература

1. Гмурман В. Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – 479 с.
2. Холлендер М. Непараметрические методы статистики / М. Холлендер, Д. Вульф. – М: Финансы и статистика, 1983. – 518с.